OBJECTIVE EVALUATIONS OF SYNTHESISED ENVIRONMENTAL SOUNDS

David Moffat *

Centre for Digital Music,
Queen Mary University of London
London, UK
d.j.moffat@qmul.ac.uk

Joshua D. Reiss

Centre for Digital Music,
Queen Mary University of London
London, UK
joshua.reiss@qmul.ac.uk

ABSTRACT

There are a range of different methods for comparing or measuring the similarity between environmental sound effects. These methods can be used as objective evaluation techniques, to evaluate the effectiveness of a sound synthesis method by assessing the similarity between synthesised sounds and recorded samples. We propose to evaluate a number of different synthesis objective evaluation metrics, by using the different distance metrics as fitness functions within a resynthesis algorithm. A recorded sample is used as a target sound, and the resynthesis is intended to produce a set of synthesis parameters that will synthesise a sound as close to the recorded sample as possible, within the restrictions of the synthesis model. The recorded samples are excerpts of selections from a sound effects library, and the results are evaluated through a subjective listening test. Results show that one of the objective function performs significantly worse than several others. Only one method had a significant and strong correlation between the user perceptual distance and the objective distance. A recommendation of an objective evaluation function for measuring similarity between synthesised environmental sounds is made.

1. INTRODUCTION

The field of sound synthesis has seen significant work in a range of areas including effective and efficient replication of existing sounds or creation of new sounds. Sound synthesis evaluation can take many different forms. Ten different evaluation criteria for evaluation of synthesis techniques were presented by [1], in which half of the criteria are based on control and parameterisation, and only two evaluation criteria relate to the sonic properties of the synthesis. One of the key aims of sound synthesis is to produce a realistic sound, with the added ability to control or interact with the sound [2, 3]. Despite this, there is limited evaluation of sound synthesis systems and their ability to produce realistic convincing sounds [4, 5].

This paper proposes a comparison of sound similarity measures, through resynthesis. The aim is to identify an objective measure that can encapsulate the perceptual similarity of sounds. Optimization of this measure would then select appropriate parameters for a synthesis engine to match a given sound, Optimisation of synthesis parameters to evaluation of sound perception has been previously demonstrated [6]. Parameter selection can be viewed as an optimisation problem in which synthesis parameters are dimensions through a fitness landscape. In many cases, we are searching through highly nonlinear search spaces, and thus evolutionary optimisation functions are effective methods to use [7, 8, 9].

* This paper is supported by EPSRC Grants EP/L019981/1 and EP/M506394/1.

Table 1: Range of Objective Evaluation Metrics used in Current sound synthesis Research

Research	Objective Evaluation Methods
[11]	Fundamental Frequency
	Spectral Centroid
	First 4 Harmonics
	Zero Crossing Rate
[12]	Spectrogram
[13]	Spectrogram
	Num and Position of Harmonics
[14]	Spectrogram
	Magnitude Spectrum
[15]	Magnitude Spectrum
[16]	MFCC vector correlation
[17]	Spectrogram envelope
[18]	Error between STFT bins
[19]	PEAQ
[20]	Least Square Error (LSE) in FD
	Simultaneous Frequency Masking (SFM)
[21]	DCT of MFCC
	Spectral Shape
	Attack and Decay Characteristics
	Duration

Section 2 will present background literature and motivate the requirement for a generalisable objective measure for synthesised sounds. The objective metrics and evaluation framework will be presented in Section 3. The subjective listening test is presented in Section 4. Results of the subjective and objective measures are given in Section 5. Recommendations for synthesis evaluation metrics are presented in Section 6, and final comments and outline of impact in the community are presented in Section 7.

2. BACKGROUND

The research aims of sound synthesis are to produce realistic and controllable systems for artificially replicating real world sounds. Current research generally focuses on either implementation efficiency, interfacing control or physical modelling, and provides very limited evaluation. There is little or no research on comparison of existing synthesis techniques [5]. Subjective evaluation is occasionally used in current sound synthesis research [4, 10], however objective evaluation is rarely used and there is no consistency in metrics that are used. A summary of sound synthesis papers that use objective evaluation is presented in Table 1. The variety of different objective measures and methods used within Table 1, shows that there is a lack of inconsistency in method for objective

evaluation.

[2] and [22] both evaluated work based on its interactivity, which often measures the parameter mapping more than the quality of the sound synthesis. Within [23], comparison of two similarity measures was performed, the MFCC distance and an audio feature vector distance. The results were evaluated with a subjective listening test. [24] objectively compares different wavetable synthesis methods using "Relative Spectral Error", with no comparison to samples or perceptual evaluation. [18] also calculated the error of bins from the Short Time Fourier Transform (STFT), between the reference and the synthesised sample. [25] utilised sound texture statistics for resynthesis work by [6] by enforcing a set of statistics on an STFT representation of an audio signal.

[21] evaluated synthesis parameter selection using a range of low level audio features, such as Fundamental Frequency, Spectral Shape, Envelope Characteristics, and Overall Duration. [21] used the DCT of the MFCCs as a sound similarity measure, to determine how similar the synthesised sound was to a recorded sample. Similarly, [16] performed correlations between MFCC vectors within adjacent frames, as a similarity measure for audio textures. [11] compared a synthesis method to recorded samples, through visual comparison of spectrograms, and comparison of some low level audio features, such as fundamental and first 4 harmonic frequencies, spectral centroid and zero crossing rate. No comparison with other synthesis methods was undertaken and no perceptual evaluation. In contrast, [26, 27] builds a physically inspired model where the physical properties measured vs. estimated are compared. The output time domain and spectrogram signals are compared visually, including locations of fundamental and harmonics. [17] used the loudness curve weighted Equivalent Rectangular Bands (ERB) envelope to perform grain selection within a granular synthesis approach. [19] attempted to evaluate the perceptual similarity of a piano note synthesis method with a sample using PEAQ, an algorithm designed for determining the quality of audio compression codecs which analyses the sound on a sample by sample basis to determine any perceptual artifacts. Where perception was considered, the notes will never be exactly the same if played with slightly different attack or at a different sample time, thus resulting in a perceptual difference where none exists.

There have been a number of approaches to searching audio parameter spaces, within a synthesised environment. An iterative process to control parameters and minimise a set of perceptually motivated audio features was developed by [6, 28]. The results were subjectively evaluated based on participants identification and synthesis realism. Further approaches using genetic algorithms have attempted to modify musical parameters based on varying fitness functions. No other method performed any formal evaluation of the synthesis results, typically reporting their final distance measure. Fitness function methods are typically calculated as distances features such as between Mel Frequency Cepstrum Coefficients (MFCCs) [9], the Discrete Cosine Transform of the MFCCs [21]. The Perceptual Evaluation of Audio Quality (PEAQ [29]) distances were measured for piano string synthesis [19], where as the distance between Least Square Error(LSE) of time domain waveform, LSE of spectrograms and LSE of spectrograms with some masking weighting were all used as distance measures [7]. [8] used sets of different audio features to measure distances.

3. OBJECTIVE MEASURE THROUGH SYNTHESIS

In this section, the methodology of evaluating a range of objective measures will be presented. The principle is that evaluation of different objective measures can be compared through resynthesis. By using the objective measure as fitness function in an iterative synthesis process, we can identify which measure best encapsulates aspects of the perception of the sounds. Every synthesised sound will be produced with the intention of sounding as close to a recorded sample as possible, and if an objective measure is able to produce this sound, then the objective measure represents the perceptual similarity of the sounds.

3.1. Sound Synthesis Methods

Four different sound effects were used for evaluation purposes. All of them are available and hosted online as part of the FXive synthesis platform [30, 31]. All synthesis methods were originally derived from [32] and are all examples of physically inspired synthesis methods, as they are commonly available open source implementations of synthesis methods.

Fire The fire synthesis model is a noise shaping synthesis method. Individual sonic components of a fire, the hiss, crackle and lapping, are all modelled though filtered and envelope shaped noise signals. Three control parameters are exposed to the user, which are *lapping*, *hissing* and *crackling*.

Rain In the rain model, components of rain are broken into a number of categories. Ambience, which is modelled as constant shaped noise, droplets, rumble and drips. Three control parameters are exposed to the user, which are *density*, *rumble* and *ambience*.

Stream The stream is modelled entirely on the bubbling sounds that are made as water runs over substances, based on control of filtered chirp sounds. Three control parameters are exposed to the user, which are *bubbles*, *frequency* and *filter Q*.

Wind The wind model uses a varying filtered noise approach, where wind parameters control the overall envelope of the sound. Different wind hitting materials, such as door or branches/wires, select the timesteps over which the wind envelope shaping will occur. Ten parameters are exposed to the user: *Wind Speed*, *Gustiness*, *Squall*, *Buildings*, *Doorways*, *Branches*, *Leaves*, *Pan*, *Directionality* and *Gain*. The parameters *Pan*, *Directionality* and *Gain* were all left constant at their default values, as discussed in Section 3.1.

Parameters Not Changed Several parameters were not used, to limit the search space and as these parameters were considered to make no immediate impact to the synthesis of the sound. During analysis, all samples were loudness normalised, so output gain controls were redundant. As no evaluation metric used spatial aspects to evaluate synthesis, pan controls were also not considered. With each sound effect, there was the ability to apply a range of audio effects, including equalisation, distortion, delay, convolution reverb and HRTF spatialisation. However, because all of these controls can be added to every single synthesised sample, we felt this would significantly grow the search space without significant improvements in the synthesis. The impact of individual audio effects on the perceived realism of a synthesised sound is out of the scope of this work.

3.2. Parameter Optimisation

The parameters of each synthesis model were optimised using particle swarm optimisation. Particle swarm optimisation is an evolutionary inspired population based optimisation technique in which a swarm of particles iteratively propagate in a search space, where a weighting between individual and global preferences are modelled. Each particle is evaluated with a fitness function, and we use this fitness function to compare each of our objective functions presented in Section 3.3. Particle swarm is an effective optimisation method for highly nonlinear search spaces, and there are many examples of evolutionary algorithms applied to audio research [7, 8, 9, 33, 34]. A comprehensive overview of particle swarm optimisation is presented in [35].

3.3. Objective Function

The fitness functions were taken from literature, and their features used for evaluation are described in Table 2. To standardise implementations, all audio features were extracted using Essentia [36, 37].

Table 2: Attributes of Each Objective Function

Objective Function	Features and Attributes
Allamanche [38]	Loudness
	Spectral Flatness
	Spectral Crest Factor
Gygi [39]	Envelope Statistics
	Pitch
	Autocorrelation Waveform Peaks
	Spectral Centroid
	Spectral Moments
	Frequency Band Energy
	Modulation Statistics
	Subband Correlation
	Spectral Flux
MFCC [9]	MFCC
Moffat [40]	Loudness
	Pitch
	MFCC
	Envelope Statistics
	Spectral Contrast
	Spectral Flux
PEAQ [29]	Signal Bandwidth
	Masking Content
	Modulation Difference
	Distortion
	Harmonic Structure
Wichern [41]	Loudness
	Spectral Centroid
	Spectral Sparsity
	Harmonicity
	Temporal Sparsity
	Transient Index (ΔMFCC)

The MFCC's as a similarity was motivated as an anchor within the experiment, as we expected this method to underperform in comparison to other objective functions.

4. SYNTHESIS EVALUATION - LISTENING TEST

4.1. Participants

19 participants took part in the experiment, of which 12 were male and 7 female. The average age 29 and standard deviation of 3. The average test duration was 23 minutes, so fatigue was not an issue. The procedure was approved by the local ethics committee.

4.2. Experimental Setup

The experiment was set up as listening test, performed in Queen Mary Studio [42], and participants auditioned sounds over a pair of high quality calibrated PMC speakers. Participant were asked to adjust the volume of the audio to a comfortable level at the beginning of the test and refrain from adjusting it. All volume adjustments were recorded during the test. The listening test was set up using the Web Audio Evaluation Tool [43]. The listening test is available[1] with the same user interface and set of samples that were used by participants.

4.3. Materials

Participants were asked to evaluate sound samples for four categories (fire, rain, stream and wind). In each category six synthesised samples were provided and compared to a recorded sample reference. All samples were 48kHz wav files, and loudness normalised in accordance with [44]. Each category had one anchor, where random parameter values were used to generate a sample. The reference samples were all selected from a professionally available sound effects library[2].

The anchors were included to encourage participants to use the entire evaluation scale, and we could review how samples were distributed within that scale, in accordance with [45]. The anchor ensures that there is a lower limit sample to compare against. It also performs as a confirmation that a participant has fully understood the requirements for the experiment. If a participant rated the anchor as higher than the sample, then we would infer that the participant may not have fully understood the requirements, or may have some hearing defect.

4.4. Procedure

Participants were provided with instructions as to the experiment they were to undertake, and were asked to provide their native spoken language, whether they had previous experience of listening tests and whether they would consider themselves as accomplished musicians or audio engineers.

Participants were then asked to rate how similar they perceived a set of given samples to a provided reference. Participants were provided with a continuous linear scale on which to rate all sounds, labeled from "most similar" to "very different". All sounds were rated on a single horizontal scale, to encourage inter-sample comparison. Participants did not have any information regarding the samples, other than that they were all synthesised and the names of the four sound classes used in the experiment. Samples started off at a randomised position on the scale. Both the ordering of categories and the initial ordering of samples within a category were randomised, to remove bias effects.

[1] `http://goo.gl/fusJv3`

[2] `https://www.prosoundeffects.com/hybrid-library/`

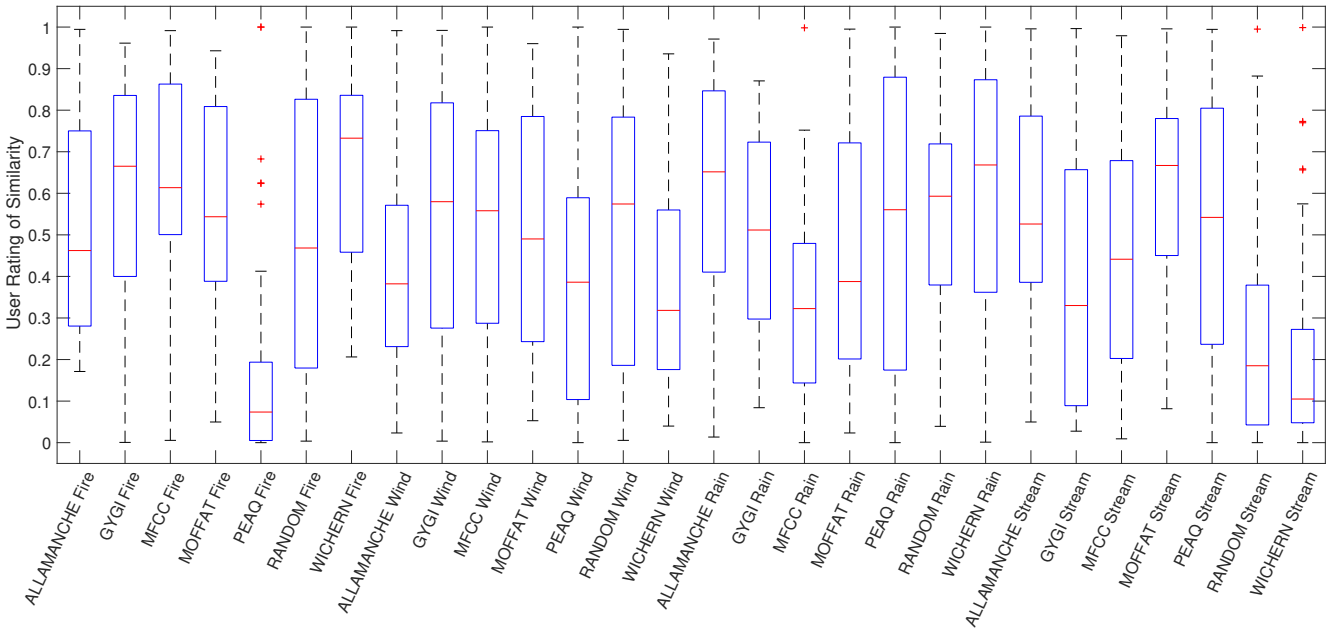

Figure 1: Distribution of User Similarity Ratings over Objective Function and Synthesis Model

5. RESULTS

One participant's results was identified as an outlier as over 30% of their answers was more than three scaled median absolute deviations from the median result. As such all results presented are of the remaining 18 participants. User similarity ratings are presented in Figure 2, where the distributions of the results can be seen.

A Shapiro-Wilk normality test showed that the data is not-normally distributed ($W = 0.95208$, $p < 2.2e\text{-}16$). A Kruskal Wallis test was performed to evaluate the impact of each objective function. A significant difference between the objective evaluation methods was found ($H=18.2$, $p=0.0057$). A post-hoc multiple comparison was performed, with results presented in Table 3.

5.1. Results per Synthesis Method

Table 3 shows that across all sound synthesis models, there is limited consistent variation. The PEAQ objective function is significantly worse than both Allamanche and Moffat. There are no further significant results at this level. To analyse the data further, we investigated the results per synthesis method, as shown in Figure 1. Kruskal Wallis tests were performed to identify the impact of each objective function for each synthesis method. The results show that there are significantly different grouping in three of the four sound synthesis methods. These results are presented in Tables 4-6. Within the wind synthesis method, no significant different in perceptual similarity to the reference sample were found between different objective synthesis methods ($H=11.72$, $p=0.069$).

As seen in Table 4, the PEAQ method is significantly worse than every other objective evaluation function with regards to fire sounds. But for rain sounds, in Table 5 MFCCs are significantly worse than Allamanche, PEAQ, random and Wicherni. For stream sounds, Table 6 shows that Allamanche, Moffat and PEAQ are all significantly better than both random and Wichern. MFCC is also significantly better than Wichern, and Moffat is significantly better than Gygi.

5.2. Comparison with Objective Function Results

Each of the objective functions also produced a distance measure, which is the value that was minimised as part of the synthesis. These distances indicate how successful the synthesis method believes it has performed in each case. The objective distances are compared with the perceptual distances, and are plotted in Figure 3, along with linear regression lines of best fit. The user similarity ratings were inverted to make the graphical representation easier to interpret, and correlations more clear. Each of the objective and subjective results were correlated, using a Spearman correlation, for non-parametric data, and the results presented in Table 7. Only the Wichern result is statistically significant, with a strong positive correlation.

6. DISCUSSION

Table 3 shows minimal significant variation in the distributions of similarity ratings. Overall Moffat performs as the best objective evaluation method, whereas Allamanche is a good options with a lower variance in the data. PEAQ performs the worst, and is significantly worse than both Allamanche and Moffat, which is the only significant generalised result.

For further analysis, we look into the breakdown per synthesis method. Within the fire sound, every objective function was significantly better than PEAQ. PEAQ is the only method that models distortion and bandwidth, and it is believed that these components of the objective function caused it to perform poorly for fire. A large portion of a fire sound is crackling and popping, and broadband noise. As PEAQ is designed for evaluation the quality of audio compression algorithms, it is designed to be sensitive to cracking and distortion artefacts. However, this is principally what makes up a fire sound. As such, it is expected that PEAQ failed to

Table 3: Multiple Comparisons Test Significance Results for All Synthesis Models, Kruskal Wallis Results (H=18.2, p=0.0057)

All Synthesis Methods	Allamanche	Gygi	MFCC	Moffat	PEAQ	Random	Wichern
Allamanche	.	o	o	o	**	o	o
Gygi	o	.	o	o	o	o	o
MFCC	o	o	.	o	o	o	o
Moffat	o	o	o	.	*	o	o
PEAQ	**	o	o	*	.	o	o
Random	o	o	o	o	o	.	o
Wichern	o	o	o	o	o	o	.

o > 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001, . = no comparison made

Table 4: Multiple Comparisons Test Significance Results for Fire Synthesis Method, Kruskal Wallis Results (H=53.19, p=1.08e-9)

Fire	Allamanche	Gygi	MFCC	Moffat	PEAQ	Random	Wichern
Allamanche	.	o	o	o	***	o	o
Gygi	o	.	o	o	****	o	o
MFCC	o	o	.	o	****	o	o
Moffat	o	o	o	.	****	o	o
PEAQ	***	****	****	****	.	***	****
Random	o	o	o	o	***	.	o
Wichern	o	o	o	o	****	o	.

o > 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001, . = no comparison made

appropriately model fire due to the wide-band, impulsive nature of the sound, which PEAQ is often identifies as a flaw. It is suspected that PEAQ will also fail to accurately model other sounds that are broadband and highly impulsive, such as applause [46] or gunshot [47] sounds.

Within the rain sounds, the MFCC evaluation metric performed significantly worse than Allamanche, PEAQ, Wichern and random. MFCCs are often used in music information retrieval as a descriptor for timbre. However, the variation in rain sounds are less timbral and more related to the ambient noise versus individual impulsive tones. The separation between constant noise tones and impulsive tones will not be identified by MFCCs. As MFCCs are no better than the random parameters, it is clear that MFCCs are not a good measure for parameter estimation within rain sounds. There is no other significant variation in objective evaluation functions. Wichern was the only method to perform better than random parameter selection, though this was not significantly better. This could be due to the random parameters being very good parameters selected by chance, or that there is limited variation within the synthesis method.

Regarding stream sounds, Figure 1 shows that Wichern and random both perform poorly, and are significantly worse than Allamanche, Moffat and PEAQ methods, and Alllamanche is significantly worse than MFCCs. It is suspected that this is due to Wichern primarily looking at harmonic content and transient sounds, where less attention was paid to broadband sound similarities. Within the stream model, most water noises will be highly broadband signals, and Wichern will most likely tend to produce more harmonic tuned sounds, than those present in a real signal. Wichern and random are not significantly worse than Gygi, which is most likely due to the large variation in the distribution of the Gygi results. This suggests that individuals were undecided or opinions were split on the result. Moffat was the best performing result and is significantly better than Gygi, along with random and Wichern. It is suspected that this is due to the inclusion of the spectral contrast feature. Spectral contrast is an audio feature that identifies the peaks and valleys in the magnitude spectrum, and performs dimensionality reduction on the result. Spectral contrast is often considered an effective method for evaluating audio masking and for identifying variations high contrast variations in frequency spectra.

The wind model failed to produce any significant difference between any objective metrics. Gygi performed the best, closely followed by random parameter allocation, but all methods are fairly similar to each other. This could be a failing of the synthesis model, as there were highly harmonic artefacts within the synthesis model, that no parameters could be removed. Further investigation of the synthesis model shows that a number of filter center frequencies are hard-coded into the model, which most likely led to inconsistent and inconclusive results. It is also possible that the number of parameters may also have influenced the results. Wind had more than twice the parameters to optimise compared to any other synthesis model, which the particle swarm algorithm may have had challenges optimising. The larger search space may have lead to issues in finding appropriate minima.

Each of the objective functions were compared and grouped in terms of how their effectiveness on a 1-5 rating scale, as presented in Table 8. It can be seen that the Gygi method performs best for both fire and wind sounds and fairly well for rain sounds, but is one of the worst objective measures for the stream sound. Gygi contains a large set of parameters relating to subband correlations and modulation statistics, which have been tied to the human auditory system [6]. As such, Gygi method seems to be the best overall performer, as consistently produced reasonable results in all cases, and between that and Moffat, it never produced the worst results. Moffat performed best overall, and was best for wind sounds, which it is suspected is due to the spectral contrast feature. It also performed reasonably well for fire and rain sounds, as the spectral contrast and spectral flux sounds will perform well for granular impulsive sounds. The Allamanche method performs best for rain sounds and reasonably well for stream sounds, but is

Table 5: Multiple Comparisons Test Significance Results for Rain Synthesis Method, Kruskal Wallis Results (H=26.81, p=1.57e-4)

Rain	Allamanche	Gygi	MFCC	Moffat	PEAQ	Random	Wichern
Allamanche	.	o	***	o	o	o	o
Gygi	o	.	o	o	o	o	o
MFCC	***	o	.	o	*	*	***
Moffat	o	o	o	.	o	o	o
PEAQ	o	o	*	o	.	o	o
Random	o	o	*	o	o	.	o
Wichern	o	o	***	o	o	o	.

o > 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001, . = no comparison made

Table 6: Multiple Comparisons Test Significance Results for Stream Synthesis Method, Kruskal Wallis Results (H=54.91, p=4.84e-10)

Stream	Allamanche	Gygi	MFCC	Moffat	PEAQ	Random	Wichern
Allamanche	.	o	o	o	o	***	****
Gygi	o	.	o	*	o	o	o
MFCC	o	o	.	o	o	o	*
Moffat	o	*	o	.	o	****	****
PEAQ	o	o	o	o	.	**	**
Random	***	o	o	****	**	.	o
Wichern	****	o	*	****	**	o	.

o > 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001, . = no comparison made

Table 7: Correlations of Objective Function Distance Measure with Mean User Similarity Rating

Objective Function	Correlations ρ	P-Value p
Allamanche	-0.3095	0.4618
Gygi	-0.0952	0.8401
MFCC	0.0238	0.9768
Moffat	-0.3095	0.4618
PEAQ	-0.4059	0.3155
Wichern	**0.7857**	**0.0279**

Table 8: Ratings of Success of each Objective Evaluation Method

	Overall	Fire	Rain	Stream	Wind
Allamanche	2	4	5	1	1
Gygi	2	1	1	3	4
MFCC	2	1	2	5	3
Moffat	1	3	3	4	1
PEAQ	5	5	5	3	2
Wichern	4	1	5	1	5

1 = Best, 5 = Worse. Ratings were created manually, based on ranking and clustering of results

one of the worse methods for wind and fire sounds. This suggests that the spectral characteristics are more complex for wind and fire sounds, as Allamanche only uses a spectral flatness and spectral crest factor as the evaluation, as all samples were loudness normalised before analysis. PEAQ performed worse overall, through performing worse in both fire and rain sounds, however performed reasonably we for stream and wind sounds. This demonstrates that PEAQ represents broadband noisy signals fairly well, however the low level textual and highly impulsive sounds are not effectively modelled by this method. The Wichern method is highly inconsistent as it performs best for fire and stream however is the worse for rain and wind sounds.

Wichern was the only objective evaluation method where the objective distance significantly correlated with the perceptual distance ratings. The correlations of the objective distance are a vital aspect of any objective evaluation function, where it is possible to predict how well the objective function performs and how effective the synthesised sound is.

7. CONCLUSION

A set of six different objective evaluation functions, for measuring similarity between environmental sounds, were tested and compared, through their ability to direct a resynthesis algorithm towards an appropriate parameter setting. In the general term, across four different types of sounds, there was no significant winner. The PEAQ method performed the worse, performing significantly worse than both Moffat and Allamanche. This demonstrates that PEAQ is not a suitable for evaluating sound similarity in a range of different cases, though it was effective for comparing broadband noisy signals, such as wind. The results demonstrate that there is currently no unilateral objective evaluation function, an consistently no method is a clear winner in most cases. One of the causes of this could be the failings or limitations of the synthesis models used. The limitation for each method to produce a wide range of sounds, could result in many different samples being challenging to synthesize, and thus cause all methods to underperform.

Despite this, the Wichern method results correlate significantly and strongly perceptual distance measures. This suggests that the

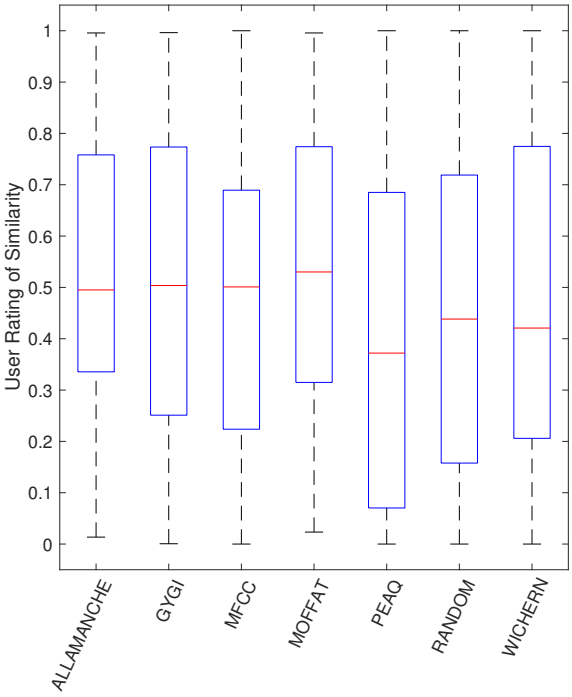

Figure 2: Distribution of User Similarity Ratings per Objective Function

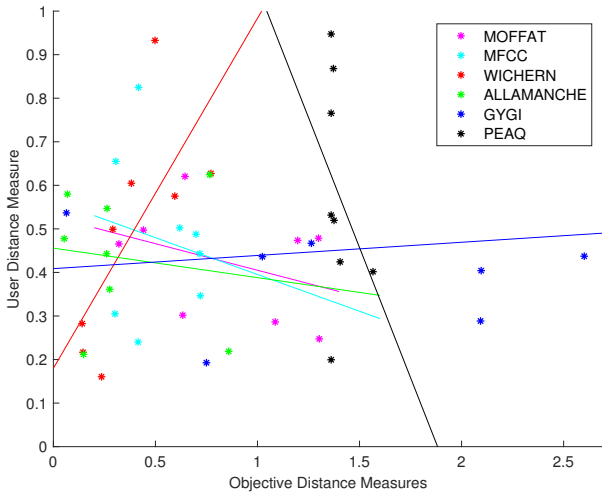

Figure 3: Inverse User Similarity Compared Against Objective Distance Metric for Each Objective Function, with Linear Best Fit Lines

Wichern method can be used as an effective distance metric, comparing similarity between different sets of sounds. Further evaluation with different synthesis methods is required to verify these results and to identify whether the synthesis methods themselves impacted the results.

The use of further different sounds samples and sound classes would also provide further data points, which would aid in correlating the objective results with the perceptual ratings. This would ensure that the results can be applied to a range of different sound types. Furthermore, there were some cases where the synthesis method was not capable of producing a very similar sample. In which case, careful improvement and selection of synthesis methods and samples could be made in future work. Further evaluation of different perceptual measures of similarity, and comparison of objective measures with expert human parameter modification could also be performed.

8. REFERENCES

[1] D. Jaffe, "Ten criteria for evaluating synthesis techniques," *Computer Music Journal*, vol. 19, no. 1, pp. 76–87, 1995.

[2] N. Böttcher and S. Serafin, "Design and evaluation of physically inspired models of sound effects in computer games," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*, London, 2009, AES.

[3] D. Moffat, R. Selfridge, and J. D. Reiss, "Sound effect synthesis and control," in *Foundations in Sound Design: an interdisciplinary approach*, Michael Filimowicz, Ed., vol. Volume 2: Interactive Media. Routledge, 2018.

[4] D. Moffat and J D. Reiss, "Perceptual evaluation of synthesized sound effects," *ACM Transactions on Applied Perception (TAP)*, vol. 15, no. 2, pp. 19, March 2018.

[5] D. Schwarz, "State of the art in sound texture synthesis," in *14th International Conference Digital Audio Effects (DAFx)*, Paris, France, 2011, pp. 221–231.

[6] J. McDermott and E. Simoncelli, "Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.

[7] R. Garcia, "Automating the design of sound synthesis techniques using evolutionary methods," in *COST G-6 Conference on Digital Audio Effects, Limerick, Ireland*, 2001.

[8] J. McDermott, N. Griffith, and M. O'Neill, "Evolutionary computation applied to sound synthesis," in *The Art of Artificial Evolution*, pp. 81–101. Springer, 2008.

[9] M. Yee-King and M. Roth, "A comparison of parametric optimization techniques for musical instrument tone matching," in *Audio Engineering Society Convention 130*, 2011.

[10] R. Selfridge, D. Moffat, and J. D. Reiss, "Real-time physical model for synthesis of sword swing sounds," in *International Conference on Sound and Music Computing (SMC)*, Espoo, Finland, July 2017.

[11] S. Hendry and J. D. Reiss, "Physical modeling and synthesis of motor noise for replication of a sound effects library," in *Audio Engineering Society Convention 129*, Los Angeles, CA, USA, 2010.

[12] M. Gasparini, P. Peretti, S. Cecchi, L. Romoli, and F. Piazza, "Real time reproduction of moving sound sources by wave field synthesis: Objective and subjective quality evaluation," in *Audio Engineering Society Convention 130*, 2011.

[13] R. Selfridge, D. Moffat, J. D. Reiss, and E. J. Avital, "Real-time physical model for an aeolian harp," in *International Congress on Sound and Vibration*, London, UK, July 2017.

[14] R. Selfridge, D. Moffat, and J. D. Reiss, "Physically derived sound synthesis model of a propeller," in *ACM Audio Mostly Conference*, London, UK, August 2017.

[15] R. Selfridge, D. Moffat, and J. D. Reiss, "Sound synthesis of objects swinging through air using physical models," *Applied Sciences*, November 2017.

[16] L. Lu, L. Wenyin, and H.-J. Zhang, "Audio textures: Theory and applications," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 2, pp. 156–167, 2004.

[17] S. O'Leary and A. Robel, "A montage approach to sound texture synthesis," in *22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, 2014, pp. 939–943.

[18] M. Athineos and D. P. W. Ellis, "Sound texture modelling with linear prediction in both time and frequency domains," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003, vol. 5, pp. 648–51.

[19] B. Hamadicharef and E. Ifeachor, "Perceptual modeling of piano tones," in *Audio Engineering Society Convention 119*, Barcelona, Spain, Oct 2005.

[20] R. A. Garcia, "Automatic generation of sound synthesis techniques," M.S. thesis, Massachusetts Institute of Technology, 2001.

[21] S. Heise, M. Hlatky, and J. Loviscach, "Automatic cloning of recorded sounds by software synthesizers," in *Audio Engineering Society Convention 127*, New York, USA, 2009.

[22] R. Nordahl, S. Serafin, and L. Turchet, "Sound synthesis and evaluation of interactive footsteps for virtual reality applications," in *IEEE Virtual Reality Conference*, Waltham, MA, USA, 2010, pp. 147–153, IEEE.

[23] D. Schwarz and S. O'Leary, "Smooth granular sound texture synthesis by control of timbral similarity," in *Sound and Music Computing (SMC)*, 2015, p. 6.

[24] A. Horner and S. Wun, "Evaluation of iterative matching for scalable wavetable synthesis," in *Audio Engineering Society Conference: 29th International Conference: Audio for Mobile and Handheld Devices*, Seoul, Korea, 2006.

[25] Wei-Hsiang Liao, Axel Roebel, and Alvin Su, "On the modeling of sound textures based on the stft representation," in *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*, 2013, p. 33.

[26] R. Selfridge, J. Reiss, E. Avital, and T. Xiaolong, "Physically derived synthesis model of an aeolian tone," in *141th Audio Engineering Society Convention*, Los Angeles, CA, USA, 2016.

[27] R. Selfridge, D. Moffat, J. D. Reiss, and E. Avital, "Creating real-time aeroacoustic sound effects using physically derived models," *Journal of the Audio Engineering Society (to appear)*, 2018.

[28] J. McDermott, D. Wrobleski, and A. Oxenham, "Recovering sound sources from embedded repetition," *Proceedings of the National Academy of Sciences*, vol. 108, no. 3, pp. 1188–1193, 2011.

[29] T. Thiede, W. Treurniet, et al., "PEAQ-The ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.

[30] P. Bahadoran, A. Benito, T. Vassallo, and J. D. Reiss, "Sound effect synthesis," 2017, UK Patent App. Num. N411552GB HHG.

[31] P. Bahadoran, A. Benito, T. Vassallo, and J. D. Reiss, "FXive: A web platform for procedural sound synthesis," in *Audio Engineering Society Convention 144*, Milan, Italy, 2018.

[32] A. Farnell, *Designing sound*, MIT Press Cambridge, UK, 2010.

[33] T. Mäkinen, S. Kiranyaz, J. Pulkkinen, and M. Gabbouj, "Evolutionary feature generation for content-based audio classification and retrieval," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. IEEE, 2012, pp. 1474–1478.

[34] D. Ronan, Z. Ma, P. Mc Namara, H. Gunes, and J. D. Reiss, "Automatic minimisation of masking in multitrack audio using subgroups," *ArXiv e-prints*, Mar. 2018.

[35] F. Marini and B. Walczak, "Particle swarm optimization (PSO). a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 153–165, 2015.

[36] Dmitry Bogdanov et al., "Essentia: An audio analysis library for music information retrieval," in *International Symposium on Music Information Retrieval (ISMIR)*, 2013, pp. 493–498.

[37] D. Moffat, D. Ronan, and J. Reiss, "An evaluation of audio feature extraction toolboxes," in *Proc. 18th International Conference on Digital Audio Effects (DAFx-15)*, November 2015.

[38] E. Allamanche, J. Herre, O. Hellmuth, et al., "Content-based identification of audio material using MPEG-7 low level description.," in *ISMIR*, 2001.

[39] B. Gygi, G. Kidd, and C. Watson, "Similarity and categorization of environmental sounds," *Perception & psychophysics*, vol. 69, no. 6, pp. 839–855, 2007.

[40] D. Moffat, D. Ronan, and J. D. Reiss, "Unsupervised taxonomy of sound effects," in *Proc. 20th International Conference on Digital Audio Effects (DAFx-17)*, Edinburgh, UK., September 2017.

[41] G. Wichern, H. Thornburg, B. Mechtley, et al., "Robust multi-features segmentation and indexing for natural sound environments," in *Content-Based Multimedia Indexing, 2007. CBMI'07. International Workshop on*. IEEE, 2007, pp. 69–76.

[42] M. Morrell, C. Harte, and J. Reiss, "Queen Mary's "Media and Arts Technology studios" audio system design," in *Audio Engineering Society Convention 130*, 2011.

[43] N. Jillings, B. De Man, D. Moffat, and J. Reiss, "Web audio evaluation tool: A browser-based listening test environment," in *Proc. Sound and Music Computing 2015*, Maynooth, Ireland, July 2015.

[44] ITU-R BS.1387-1, "BS. 1387, method for objective measurements of perceived audio quality," Tech. Rep., ITU-R, 1998.

[45] ITU-R BS.1534-3, "BS. 1534, method for subjective assessment of intermediate quality level of audio systems," Tech. Rep., ITU-R, 2015.

[46] A. Adami, A. Taghipour, and J. Herre, "On similarity and density of applause sounds," *Journal of the Audio Engineering Society*, vol. 65, no. 11, pp. 897–913, 2017.

[47] L. Mengual, D. Moffat, and J. D. Reiss, "Modal synthesis of weapon sounds," in *61st Audio Engineering Society International Conference: Audio for Games*, 2016.

[48] Marios Athineos and Daniel PW Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5237–5245, 2007.